DOCUMENT RESUME

ED 218 317 TM 820 357

AUTHOR Quellmalz, Edys
TITLE Problems in Stabilizing the Judgment Process.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
REPORT NO CSE-R-136
PUB DATE 80
NOTE 27p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Measurement Techniques; *Scoring; *Testing Problems; Test Reliability; Validity; Writing (Composition); *Writing Evaluation; *Writing Skills

ABSTRACT

Measurement problems which jeopardize the reliability and validity of competency-based writing assessments are analyzed. Methods to stabilize rating criteria and readers' application of them are necessary. Most writing assessment programs use guidelines from norm-referenced test methodology. Use of this method of criteria application based on ranking within-occasion endangers or precludes between-occasion uniformity. Judgment stability within a session and across sessions are other problems of measurement. The instability of ratings has been a major weakness of writing skills measures. Two indicators of rating variability are discussed. Rater drift is the rater's progressive deviation within a scoring session from previously shared criteria. Scale instability is the differential application of criteria by raters in different scoring sessions. During scale development and validation, assessments should collect separate ratings on component text features that comprise a total score. Rating methods should intersperse periodic checks. Frequency of checks and nature of feedback on scoring accuracy are important. Sound writing assessment requires scale stability. (Author/DWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED218317

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

PROBLEMS IN STABILIZING THE JUDGMENT PROCESS

Edys Quellmalz

CSE Report No. 136
1980

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G. Cerag

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 820 357

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California, Los Angeles

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

Abstract

Problems in Stablizing the Judgment Process

This article analyzes a series of measurement problems that jeopardize the reliability and validity of competency-based writing assessments. The paper distinguishes between two indicators of rating variability 1) rater drift -- rater's progressive deviation within a scoring session from previously shared criteria; and 2) scale instability -- differential application of criteria by raters in different scoring sessions. Examples from research illustrate the nature and magnitude of rating fluctuations. Promising techniques are described for stabilizing raters' judgments and documenting scale stability.

Problems in Stablizing the Judgment Process

Edys Quellmalz

Center for the Study of Evaluation
University of California, Los Angeles

The increasing demand for competency assessments of complex human performance has led to renewed scrutiny of the conceptual and technical quality of prevailing testing practice. Particularly in the area of language production, i.e., writing, oral language and oral reading, researchers and practitioners assert that competency tests must provide tasks that match performance objectives and that activate cognitive processing strategies required by production rather than recognition tasks. The validity of indirect (i.e., multiple choice) measures is no longer logically, psychologically or ecologically acceptable to the majority of professionals in writing instruction and evaluation. Life is not a multiple choice. Students' language production skills, in particular, must be sufficiently proficient for students to function autonomously in the real world.

Although collecting samples of complex performance can presumably provide "direct," valid measures of content, the renowned unreliability of judging constructed responses continues to plague assessment methodology. Because direct performance samples are mediated by highly variable judgments of raters who score or characterize performance samples along some dimensions, a critical goal for performance judgment in gen-

eral, and for writing judgment in particular, is to find ways to assure that judges apply scoring criteria accurately and fairly. As a part of a broader program studying issues in test design, we have investigated dimensions of the test tasks, context and scoring that will reduce irrelevant variability in examinee and rater behavior.

This paper analyzes a series of measurement problems that jeopardize the validity of the judgment process and examines the effectiveness of methods currently employed to address these problems. Reviews of prevailing rating practices, in conjunction with cumulative empirical evidence on factors influencing judgments in domain-referenced assessment, demonstrate that direct writing assessment faces a dual validity requirement. Both the test task and the scoring procedure must meet separate conceptual and statistical validity standards. The paper elaborates the requirements for accurate and fair writing competence assessment and illustrates how state-of-the-art rating processes pose serious threats to the validity of the writing assessments.

Domain-Referenced Scoring Requirements

The avowed intent and structure of competency or domain-referenced tests require explicit, replicable scoring criteria and procedures; thus, the need for methods to stabilize rating criteria and readers' application of them is immediate and real. Soon the uniform application of performance criteria may become a legal requirement when decisions based on these tests result in life-altering consequences for students. Mandates proliferate at state and local levels for writing assessment at all levels of public

school, and large numbers of writing samples must be scored by great numbers of raters. Many assessment programs are required to provide students repeated opportunities to pass comparable forms of a test. Also built into many assessment programs is a requirement to administer comparable tests, at regular intervals, at geographically separate sites.

The purpose of these competency assessments is to monitor development of students' skills at points specified throughout their schooling, to detect skills for which they might need remedial assistance and to document skill development. A student who fails to demonstrate competency in writing, receives additional instruction, and is then retested should be judged according to the same standards at each test administration. His or her score should not depend on either the performance of a new cohort of examinees nor upon the idiosyncratic values of differently oriented sets of raters.

Unfortunately, many writing assessment programs derive their guidelines from norm-referenced test methodology. In practice, norm-referenced writing tests are scored by ranking papers within the limits of a particular sample. Essays are usually scored holistically, on generally described criteria, and involve scoring procedures where raters rank essays by sorting them into piles anchored by the range of quality of that particular sample (Conlan, 1976). Thus a particular paper's rank and/or score could change from sample to sample, if the range of the quality of the competition varied from one test group to the next. Such practices result in a "sliding scale" where the rated quality of a particular paper changes according to the quality range of papers in the group. For example,

a student might take a writing competency test in the fall, when all students, low achievers to college preparatory students, participate. A student's rank in this wide quality range is below mastery. In the spring, the student, along with the restricted range of students who failed the first administration, takes another writing competency test and just passes. Does s/he pass because intervening writing instruction has strengthened weak writing skills, or because her or his rank is higher in the restricted range of poorer writers? Present holistic scoring procedures can not provide an answer to this question. The holistic score provides no evidence of the developmental level of specific writing weakness that were low and may have improved. Despite the use of "anchor" papers during training to illustrate what a "6" or "3" had been for other groups, the most prevalent holistic scoring procedures still require raters to distribute papers across the score range.

A major measurement problem confronting many competency based writing assessments, then, is the failure to deal with the need to assure comparability of scoring <u>between</u> test occasions as well as <u>within</u> a scoring session. Such comparability would require <u>not</u> just statistical indices of rater agreement but comparisons of mean scores, since ratings within a session might agree but differ between sessions. Adopting a norm-referenced method of criteria application based on ranking within occasion imperils, if not precludes between-occasion uniformity of criteria application. Therefore two measurement problems inhere in judgment stability, stability within a session and stability across sessions.

To document scale stability, an assessment would have to intersperse anchor papers scored in previous assessments among papers rated within an on-going rating session and report comparability of anchor paper scores across test occasions and rater groups. Such documentation of comparability is conspicuously absent in both research and practice.

Research on Rating Variability

Evidence pointing to the sources and manifestations of scale instability can be found in the rapidly accumulating body of research on issues of rating variability. The instability of ratings has been a major, and generally acknowledged, weakness of measures of writing skill (Coffman, 1971b). Braddock, Lloyd-Jones and Schoer (1963) classified four sources of error: 1) the writer, 2) the assignment, 3) the rater, 4) between raters. Although considerable research within the framework of domain-referenced testing has examined dimensions of the test task that influence writer performance such as discourse aim and topic modality (Pitts, 1978; Spooner-Smith, 1978; Quellmalz, 1979; Praeter & Padia, 1980; Crowhurst, 1980), less attention has been given to the factors involved in rater behavior.

In the broadest sense, inter- and intra-rater variability are a matter of fluctuating standards of judgment. Research has amply demonstrated that anarchical scoring of essays, where raters apply their individual standards, results in high disagreement among raters from different occupations (Diederich, French, & Carlton, 1961) and even among English professors (Findlayson, 1951; McColly, 1970). Follman and Anderson (1967) demonstrated that the more homogeneous the background of raters, the more

their scoring agreed. Long ago, Eels (1930) demonstrated the problem of intra-rater criteria bias when he found that the variability in essay scores assigned even by the same reader on different occasions approached the degree of variability of scores assigned by different readers. Recognizing the magnitude of error occurring in unstructured scoring, researchers attempted to devise various techniques for controlling score variability.

Methods for Controlling Scoring Variability

The first and most critical step in stabilizing the bases of readers' judgments is to establish common, explicit scoring criteria. Criteria may either be specified deductively by invoking standards derived from the rhetorical tradition (e.g., Kinneavy, 1971) or inductively by seeking commonality among readers' comments on papers (Diederich, 1974; Freedman, 1978). Systematic training on common scoring criteria has proved to reduce some kinds of interrater variability effectively. (Stalnaker, 1934; Diederich, 1974). As a result of these pioneering studies, standard methodology now includes training of raters on the use of rating scales until a high level of agreement among raters is achieved. In a recent study of the discriminative validity of alternative scoring rubrics, Winters (1978) suggested that high rater reliability coefficients in pilot or in final rating sessions might not necessarily signal standard, uniform interpretation of rating scales over rating occasions and across rater groups. During rater training she observed that less operationalized scale rubrics stimulated extensive discussion and interpretation and suggested that different rater groups might achieve high reliability, but

have interpreted vague criteria differently by devising different specific decision rules for the same ambiguous criteria. Thus, high reliability coefficients might be obtained, but at the cost of accurate, replicable scoring. As Winters implies, redefinition of criteria by the social rating group can have serious implications for the fairness of ratings across rater groups.

Rater Drift

Even with training for rater consensus, when raters practice applying explicit criteria, rating fluctuation may still occur. The deviation of raters from previously-shared criteria is termed "rater drift" and may be signaled by lowered inter-rater reliability and differences between raters' criteria interpretation and expert-generated criterion-based ratings.

Rater drift is particularly a problem when there are large sets of papers to be scored. Shifting criteria or drift may be caused by rater fatigue, or by more systematic influences, such as the quality range of the sample of papers being read or idiosyncratically valued criteria. In a description of the rater as a source of error, Braddock et al. (1963), discussed the need for controlling for rater fatigue. They cited fatigue as a cause for raters to become severe or erratic in their evaluation or to place more weight on particularly noticeable essay elements such as mechanics. Godshalk, Swineford, and Coffman (1966) found significant differences between papers scored holistically early and later in a set of 646 papers. Coffman (1971b) warned that even when two sets of scores

derive from changing combinations of raters, "there may still be differences in the means and standard deviations attributable to order effects -- that is, the tendency of groups of raters to shift their standards as the reading proceeds" (p. 276). Coffman (1971a) also discussed raters' tendency to regress to their own internalized set of standards and recommended practice on common criteria.

Rater drift impairs the technical quality of rating results by reducing inter- and intra-rater reliability, and more importantly, compromises the validity of ratings. However, writing assessment programs do not seem to acknowledge rater drift as a validity problem, nor do they deal with rater drift directly.

State-of-the-Art Procedures for Treating Scoring Variability

Current rating procedures (Conlan, 1976; Office of the Los Angeles Superintendent of Schools, 1977) generally follow methods recommended by Braddock et al. (1963), and Coffman (1971a) and have evolved a number of methods to deal with rater variability. Typically, raters begin by practicing applying a rubric to a sample set of papers. The nature and relative specificity of scale criteria and scoring formats (holistic vs. analytic) vary, as do the weights of component criteria. Before independent rating begins, trainers conduct a reliability check. Sometimes consensus is checked statistically; sometimes it is indicated by a show of hands.

During independent ratings, methods for dealing with rater agreement tend to take two tacks: correction and maintenance. Procedures which emphasize correction use post hoc methods to treat score discrepancies.

Common options are: 1) having a third reader score any paper where the first readers disagree by more than one point; 2) using the sum of two ratings as a total score; 3) randomizing the order in which two raters score an essay in order to distribute rater error, although often the randomization occurs in a single day. These post hoc correction procedures sidestep the validity problem of the changing criteria employed by the drifting rater.

A second set of procedures for dealing with rating variability aims at maintenance of scoring accuracy. Periodic consensus checks on identical papers are interspersed at varying intervals. Checks may be common to all raters, discussed in the group, discussed within rater pairs or discussed with a "master" rater. In the procedure, discrepancies are called to the rater's attention and their bases revised. These maintenance procedures at least attempt to prevent, detect, and control scoring error by providing feedback to individual raters regarding the accuracy and consistency of their scoring decision rules.

Rating Variability in Competency Assessment Research

In a series of studies examining dimensions important in the formulation of valid, instructionally sensitive writing assessments, we documented the effects of several stringent procedures for attaining and maintaining rater congruence and fidelity to the rating scale. One component of the methodology was to develop analytic scoring rubrics referenced to basic structural features of a discourse mode. Explicit criteria were designed to reference operational, instructionally manipulatable elements

of the paper. Raters practiced applying the scoring rubric in intensive training sessions and reliability checks using generalizability statistics were calculated to assure inter-rater reliability. During final, independent ratings, common checks occurred at frequent intervals. Discrepancy resolution procedures were of several types, including group discussion or pair discussion. The research focus of these studies was on variations of the tasks of writing rather than on variables influencing the rating process, yet the accumulating data indicated that stabilizing the judgment process was a complex issue--one deserving direct experimental investigation. This conclusion derived primarily from three of our studies in which we observed rater drift surface as a problem, despite the different procedures used to prevent it. We also began to inspect indices of scale stability by looking at scores given by raters trained at different times to the same set of papers.

Rater Drift

In our writing assessment research our initial scoring concerns were to establish and maintain rater agreement. To determine that this occurred, we compared reliabilities obtained immediately after training (on a pilot test of independent ratings) and after the final ratings. Table 1 presents a comparison of generalizability coefficients marking rater agreement levels on pilot and final ratings.

- - - - - - - - - - - -
Insert Table 1 here
- - - - - - - - - - - -

The first rating procedure was employed in Study 1 where Spooner-Smith

Table 1

Comparison of Generalizability Coefficients for Rater Agreement Immediately After Training and After Final Ratings

Study 1 - Expository Scale I (Spooner-Smith, 1978)

	F GC	Dev GC	O GC	Su GC	Pa GC	M GC	Total GC
Pilot - 4 raters n=15	.94	.92	.94	.83	.94	.80	.90
Final - 2 ratings n = 112	.84	.80	.85	.85	.80	.95	.90

Study 2 - Expository Scale II (Quellmalz and Capell, 1979)

	GI GC	F GC	O GC	S GC	M GC	Total
Pilot - 4 raters	.74	.63	.74	.77	.73	
Final - 2 ratings	.67	.59	.61	.57	.52	.56

Narrative Scale II

	GI GC	F GC	O GC	S GC	M GC	Total
Pilot - 4 raters	.86	.76	.79	.76	.52	
Final - 2 ratings	.84	.60	.72	.72	.69	.83

Study 3 - Expository Scale III (Baker and Quellmalz, 1980)

	GI GC	Gen Comp GC	Coh GC	Po GC	Su GC	M GC	Total GC
Pilot - 3 raters	.74	.65	.86	.93	.84	.71	.89
Final - 2 ratings	.66	.71	.62	.83	.71	.76	.81

Narrative Scale III

	GI GC	Gen Comp GC	Coh GC	Po GC	Su GC	M GC	Total GC
Pilot - 3 raters	.83	.75	.62	.87	.54	.85	.79
Final - 2 ratings	.70	.76	.53	.87	.67	.68	.81

KEY

GC = Generalizability Coefficient

Study 1 (Spooner-Smith, 1978)

- F = Focus
- Dev = Development
- O = Organization
- Su = Support
- Pa = Paragraphing
- M = Mechanics
- T = Total

Study 2 (Quellmalz and Capell, 1979)

- GI = General Impression
- F = Focus
- O = Organization
- Su = Support
- M = Mechanics
- T = Total

Study 3 (Baker and Quellmalz, 1980)

- GI = General Impression
- Gen Comp = General Competency
- Coh = Coherence
- Po = Paragraph Organization
- Su = Support
- M = Mechanics
- T = Total

(1978) compared direct and indirect measures of writing competence. Four raters received five hours practice applying an analytic rubric, Expository Scale I, to a set of papers representative of the experimental set. The top table presents Spooner-Smith's interrater reliabilities for four raters on the pilot test conducted immediately after training and on the final independent ratings of the experimental papers. During the final independent scoring, raters read, rated and discussed discrepancies on a common paper as a group approximately every hour to check adherence to criteria. While the total score reliability on the final ratings remained high, reliabilities of four of the six subscales dropped as much as .14, indicating some degree of rater drift from original consensus levels.

The second rating procedure occurred in Study 2 (Quellmalz & Capell, 1979) which compared writing performance in different discourse and response modes. Following scale training procedures employed by Spooner-Smith (1978), pilot tests of interrater reliabilities for two revised analytic rubrics, Expository Scale II and Narrative Scale II, checked level of agreement of the four raters prior to final rating. Additional training occurred on any subscale where the generalizability coefficient was less than .70. During final scoring, rater pairs read and discussed common papers after every 20 independent ratings. The two tables for Study 2 indicate, again, that agreement levels on the total scores were acceptably high, but that reliabilities on three of the expository subscales deteriorated as much as −.20. The interpretation of these data was that the frequency and nature of the common check procedures were still not curbing rater drift adequately.

Consequently, Study 3 implemented a revised rating procedure. Study 3 (Baker & Quellmalz, 1980) investigated the effect of modality of topic presentation on eighth grade writing performance. Three raters participated in scale training for analytic Expository Scale III and Narrative Scale III. Following a pilot test of inter-rater reliability, the three raters independently scored the experimental papers. Each paper received two ratings. Common checks occurred every hour and were discussed by the entire group.

As the two tables for Study 3 indicate, agreement levels fall on General Impression, but not on the General Competency rating. Reliabilities plummetted on the expository Coherence ratings and on the Mechanics ratings of the narrative scale. These comparisons of pilot and final reliabilities for Study 3 suggested that the revised checking procedure was generally maintaining rater agreement, but still did not prevent drift on some subscales.

In a more detailed inspection of the emergence of rater drift in Study 3, we also compared reliabilities and mean scores on papers scored early and late in the rating sequence (see Table 2). Table 2 presents the early vs. late comparisons for Expository Scale III and Narrative Scale III. On the expository scale, reliabilities across all rater pairs remain high (α .76 to .85) except on the General Impression and Coherence subscales. Parametric comparisons of mean scores on early vs. late papers did not reach statistical significance, but late scored papers received slightly higher ratings than early scored papers.

Reliabilities on Narrative Scale III remained high on General Compe-

TABLE 2

Comparison of Early vs. Late Scored Papers in Study 3

(Baker and Quellmalz, 1980)

	Inter-rater Reliabilities		Mean Scores		
	Early	Late	Early	Late	t
Expository Scale III					
General Impression	α .85	.69	$\bar{X}$ 2.28 S.D. 1.07	2.29 .85	.97
General Competency	α .75	.77	$\bar{X}$ 2.20 S.D. .91	2.43 .86	.23
Coherence	α .78	.57	$\bar{X}$ 2.39 S.D. .88	2.63 .90	.21
Paragraph Organization	α .87	.86	$\bar{X}$ 2.03 S.D. 1.05	2.22 1.08	.40
Support	α .78	.76	$\bar{X}$ 2.99 S.D. .85	3.11 .90	.51
Mechanics	α .67	.82	$\bar{X}$ 2.18 S.D. .85	2.99 .76	-1.08
Total	α .87	.85	$\bar{X}$ 14.78 S.D. 4.86	15.89 4.49	-1.06
	n=40	n=40	n=40	n=40	
Narrative Scale III					
General Impression	α .78	.71	$\bar{X}$ 2.62 S.D. .92	2.19 .73	2.31*
General Competence	α .81	.78	$\bar{X}$ 2.54 S.D. .87	2.20 .78	1.84
Coherence	α .77	.46	$\bar{X}$ 2.60 S.D. .99	2.31 .59	1.60
Paragraph Organization	α .93	.85	$\bar{X}$ 2.22 S.D. 1.29	2.03 1.00	.74
Support	α .84	.84	$\bar{X}$ 2.82 S.D. .97	2.51 .68	1.68
Mechanics	α .68	.80	$\bar{X}$ 2.30 S.D. .80	2.16 .74	.82
Total	α .90	.86	$\bar{X}$ 14.35 S.D. 4.94	13.03 3.44	1.49
	n=40	n=50	n=40	n=50	

α = alpha coefficient

* $p < .05$

tence, Support, Mechanics and Total score. General Impression reliability dropped .08, Coherence dropped substantially (α .77 to .46) and Paragraph Organization fell (α .93 to .85). Contrasts of mean differences between early and late scored narrative papers revealed a significant difference on General Impression ratings. Papers scored later received lower ratings than those scored earlier. All subscale scores were lower for late scored papers. These findings are consistent with other research (Godshalk et al., 1966) that reported raters became more severe as scoring progressed. In Study 3, Expository papers were scored before Narrative papers, so late scored Narrative papers were at the very end of the entire scoring sequence.

Inspection of the scoring data from the three studies suggests that rater drift within a scoring session can occur and weaken scoring rigor. Raters' judgments waivered on some subscales more than others, signalling a need for more careful explication of criteria on those subscales and practice on their application. Since state-of-the-art procedures for controlling rater drift were employed and even refined in these studies, the data implied the need to continue to examine methodologies for detecting and preventing rater drift.

Scale Stability

A validity concern coordinate with maintenance of scale fidelity within rating occasion is assurance of judgment accuracy across rating occasions. Standards of fairness and methodological rigor mandate that criteria apply uniformly across sets of raters and sets of papers.

Prevailing practice does not seem to recognize stability as a technical problem. Large scale assessments do not routinely report and inspect a

series of rater reliabilities for separate scoring sessions. Even reliability indices are not sufficient, however. Comparisons of mean scores on common papers should supplement reliability statistics. Scale stability could be demonstrated by comparing scores on a common set of papers given by different rater sets trained separately, or by comparing scores from the same raters rating at different occasions. While we have not yet investigated this phenomenon within an experimental paradigm, we have, however, inspected scoring data gathered during the process of our other writing assessment research in an attempt to understand the nature of variables influencing scale stability.

Our Table 3 presents the means and standard deviations of essay scores given by two different rater sets to the same papers. Raters A and B scored 30 expository essays. Rater pairs 1, 2 and 3 rated these same 30 essays in the course of Study 3. Rater pairs 1, 2 and 3 were using Expository Scale III, a revision of the analytic expository rating scale used by Raters A and B. Therefore only scores from those subscales that were not significantly changed were entered into the analysis. Agreement levels were not calculated due to the small sample size.

Inspection of the means reveals that Raters A and B gave generally higher ratings than Rater pairs 1, 2 and 3. Comparisons of means for each subscale and the total score were all significant. While the small number of papers clearly limit interpretation of these data, they do document that criteria definition and application did change from one rating session to the next.

Table 3

Comparison of Essay Scores* Given by Different Rater Sets on Separate Occasions

Subscales		Ratings: Occasion 1 Raters A and B	Ratings: Occasion 2 Raters 1-6	t	df
General Competence	$\bar{X}$	2.92	1.65	2.77*	57
	s.d.	.62	.38		
	n	29	30		
Paragraph Organization	$\bar{X}$	2.19	1.46	3.67*	58
	s.d.	.98	.50		
	n	29	31		
Support	$\bar{X}$	2.76	2.07	3.25*	59
	s.d.	1.08	.50		
	n	29	32		
Total	$\bar{X}$	11.81	8.97	4.38*	59
	s.d.	3.17	1.76		
	n	29	32		

*Scores by rater pairs 1-6 were transformed from a score range of 1-6 to 1-4 to permit analyses.

In addition to looking at the scores different raters trained at separate occasions gave to the same set of papers we inspected intra-rater agreement of scores a pair of raters gave to common papers scored at different sessions. Table 4 displays means and standard deviations of a rater pair (N) which participated in two different rating sessions.

- - - - - - - - - - -

Insert Table 4 here

- - - - - - - - - - -

In Study I, rater pairs M and N scored essays from a general high school population which were then "salted in" a set of college admission essays read for Study 2. In Study 2, pair N read the eight essays they had scored previously in Study 1 and 8 additional essays from that study that they had not personally scored. The means of pair N in the two studies are fairly comparable except on Support and Mechanics. In contrast, the means of pairs M and O are substantially different. Pair O means are consistently lower. The greater stability of means for pair N may suggest that they were applying criteria in a uniform manner. Pair O was probably influenced by the overall higher quality of the college admissions sample, thus making the "salted in" general population high school seem worse. Methods for eliminating this subtle "norming" of presumably explicit criteria to the quality range of particular sample is a phenomena requiring further research.

Our intent in inspecting these admittedly limited data was to illustrate one method for tracking the stability of rating scale application. Writing assessments could systematically include a "check" set of papers in each rating session to document the comparability of judges' decision

TABLE 4

Comparison of Rater Pair Scores Across Studies

		Study 1		Study 2	
Rater Pair		M	N	N	O
CSE Subscale					
General Impression	$\bar{X}$	1.92	1.28	1.00	.94
	s.d.	1.32	1.37	1.13	.91
	n	6	8	16	16
Focus	$\bar{X}$	2.08	1.71	1.69	1.53
	s.d	.38	.9	.48	.50
	n	6	8	16	16
Organization	$\bar{X}$	2.33	1.65	1.72	1.38
	s.d..	.98	.6	86	.50
	n	6	8	16	16
Support	$\bar{X}$	2.42	2.76	2.00	1.63
	s.d.	.92	1.15	1.78	.50
	n	6	8	16	16
Mechanics	$\bar{X}$	2.50	2.20	1.78	1.75
	s.d.	.84	.70	.77	.58
	n	6	8	16	16
Total	$\bar{X}$	11.25	9.60	8.19	7.21
	s.d.	3.71	2.79	3.40	2.53
	n	6	8	16	16

rules at different rating sessions. We believe that scale stability across topics, quality range of papers and sets of raters can be achieved and that the factors influencing scale stability require systematic investigation.

Summary and Recommendations

The need for stabilizing the scoring process is critical to the validity of writing assessments. Direct evidence of student writing competence, actual written production, is a necessary condition for content and construct validity; it is not sufficient, however. Rater's judgments must be replicable and defensible. We believe that explicit rating criteria are a condition for defensibility and replicability. Our rater drift comparisons suggest that total scores and a holistic score seem to mask fluctuations in judgments on the elements that contribute to the more global summary scores. We suspect that, at least during scale development and validation, assessments should collect separate ratings on component text features such as Support and Coherence that contribute to a total score. Otherwise, there is no way to identify and track consistency of the bases for global judgments.

Certainly, scale training and an initial reliability check is essential. Rather than relying primarily on randomization or statistical procedures to correct for rater drift post hoc, rating methods should intersperse periodic checks into lengthy, independent scoring. The variables making these checks effective for maintaining agreement and scale fidelity require further investigation. Frequency of checks is one important factor; the nature of feedback on scoring accuracy is even more essential. We

are currently conducting research on methods for curbing rater drift.

Scale stability is a critical validity issue for competency-based writing assessment. Large scale assessments can, at least, document stability by tracking scoring of a core set of papers by different groups of raters. Methodologies for selecting and preventing scale instability should also receive direct experimental attention. Fair, informative, generalizable, defensible scoring procedures are necessary requirements of sound writing assessment.

References

Baker, E. L., & Quellmalz, E. S. Issues in eliciting writing performance: Problems in alternative prompting strategies. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, April 1980.

Braddock, R., Lloyd-Jones, R., & Schoer, J. Research in written composition. Urbana, Ill.: National Council of Teachers of English, 1963.

Coffman, W. E. Essay Examinations. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D. C.: American Council of Education, 1971a.

Coffman, W. E. On the reliability of ratings of essay examinations in English. Research in the Teaching of English, Vol. 5(1), Spring 1971b.

Conlan, G. How the essay in the CEEB English test is scored. Princeton, N. J.: Educational Testing Service, 1976.

Crowhurst, M. Syntactic complexity in narration and argument at three grade levels. Canadian Journal of Education, 1980.

Diederich, P. B., French, J. W., & Carlton, S. Factors in judgments of writing ability. Princeton, New Jersey: Educational Testing Service, 1961.

Diederich, P. B. Measuring growth in English. Urbana, Ill.: National Council of Teachers of English, 1974.

Eels, W. C. Reliability of repeated grading of essay-type examinations. Journal of Educational Psychology, 1930, 21.

Findlayson, D. S. The reliability of the marking of essays. British Journal of Educational Psychology, 1951, 21, 126-134.

Follman, J. C., & Anderson, J. A. An investigation of the reliability of five procedures for grading English themes. Research in the Teaching of English, 1967, 190-200.

Freedman, S. How characteristics of student essays influence teachers' evaluation. Journal of Educational Psychology, 1978, 70.

Godshalk, F. E., Swineford, F., & Coffman, W. E. The measurement of writing ability. New York: College Entrance Examination Board, 1966.

Kinneavy, J. R. A theory of discourse. In the Aims of Discourse. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1971.

McColly, W. What does educational research say about the judging of writing ability? The Journal of Educational Research, 64, No. 4, December 1970.

Office of the Los Angeles County Superintendent of Schools. A common ground for assessing competencies in written expression, review copy. Los Angeles: Division of Curriculum and Instructional Services, 1977.

Pitts, M. The relationship of classroom instructional characteristics and writing in the descriptive/narrative mode. Report to the National Institute of Education, Los Angeles: UCLA Center for the Study of Evaluation, 1978. (Grant No. OB-NIE-G-78-0213)

Prater, D., & Padia, W. Effects of modes of discourse in writing performance in grades four and six. Paper presented at the annual meeting of the American Educational Research Association, Boston, 1980.

Quellmalz, E. S. Interim report. Defining writing domains: Effects of discourse and response mode. Center for the Study of Evaluation, University of California, Los Angeles, 1979.

Quellmalz, E. S., & Capell, F. Defining writing domains: Effects of discourse and response mode. Report to the National Institute of Education, November, 1979. (Grant No. OB-NIE-G-78-0213 to the Center for the Study of Evaluation)

Spooner-Smith, L. Investigation of writing assessment strategies. Report to the National Institute of Education, November 1978. (Grant No. OB-NIE-G-78-0213 to the Center for the Study of Evaluation.)

Stalnaker, J. The construction and results of a twelve-hour test in English composition. School and Society, 1934, 39.

Winters, L. The effects of differing response criteria on the assessment of writing competence. Grant No. OB-NIE-G-78-0213, Los Angeles, California: Center for the Study of Evaluation, November 1978.